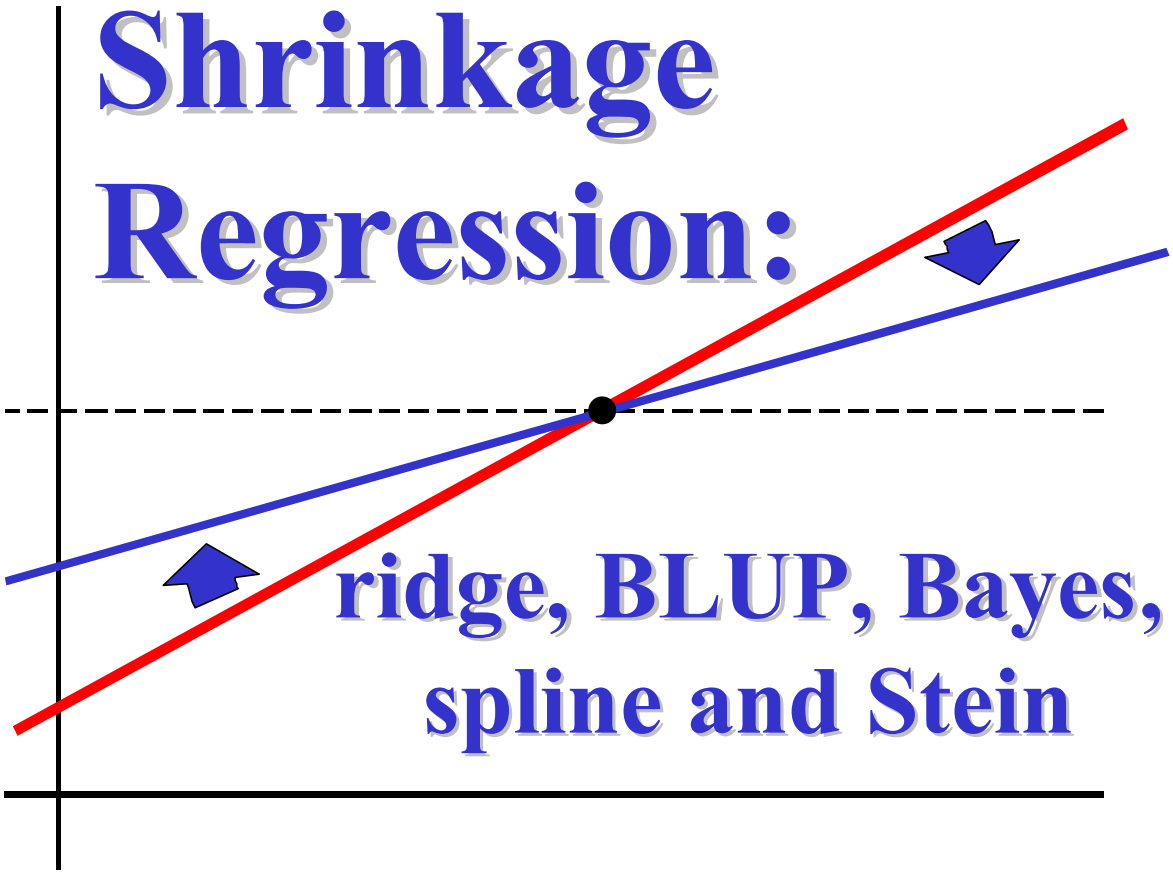



# Chapter 03: Shrinkage Regression Fundamentals

**Bob Obenchain, Ph.D.**
**softR$_X$ freeware**
**13212 Griffin Run**
**Carmel, Indiana 46033-8835**

# Chapter 3: SHRINKAGE REGRESSION FUNDAMENTALS

The regression estimators of main interest to our exposition here are known as generalized shrinkage estimators or generalized ridge regression estimators. The vector of estimators for all P of the elements of the $\beta$ coefficient vector in a linear model, such as that in equations { 2.1 } and { 2.2 } of Chapter 2, will be denoted here by the symbol $b^{\bigstar}$ and will be of the general form

$$b^{\bigstar} \;=\; G\,\Delta\,c \;\;=\; \sum_{j=1}^{j=R} \vec{g}_j \cdot \delta_j \cdot c_j \;. \qquad \{\, 3.1 \,\}$$

In equation { 3.1 }, $\vec{g}_j$ is the j-th column of the principal-axis direction-cosines matrix, G ; $\delta_j$ is the j-th diagonal element of the shrinkage factors matrix, $\Delta$ ; $c_j$ is the j-th element of the uncorrelated components vector, c , of equation { 2.16 } ; and R is the rank of the centered regressor X matrix. We will usually restrict the RANGE of interest for all R of the shrinkage factors, $\delta_1 \,, \ldots, \delta_R$, to...

$$0 \;\leq\; \delta_j \;\leq\; 1 \;. \qquad \{\, 3.2 \,\}$$

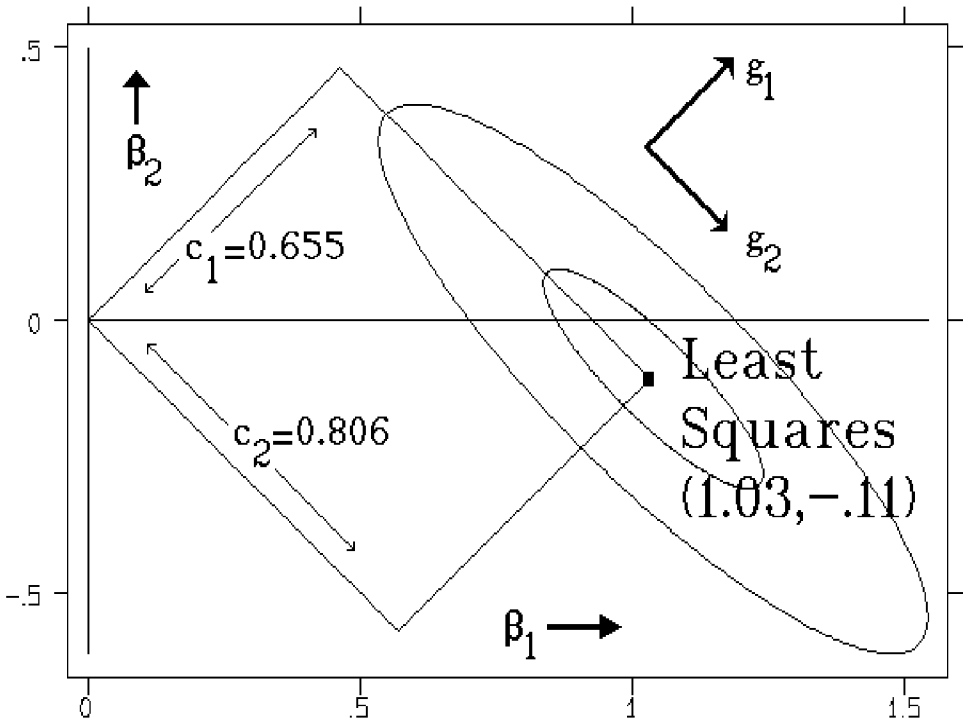
Kronecker's delta function, which will be familiar to many readers, takes on only two possible values, zero and one. In our exposition, shrinkage $\delta$ factors can take on all numerical values between zero and one, inclusive. But our use of this $\delta$ notation may help readers remember the conventional extreme values, 0 and 1, of the range allowed for shrinkage factors.

The generalized shrinkage estimator corresponding to $\delta_1 = \ldots = \delta_R = 1$ ( i.e. $\Delta = I$ ) coincides with the ordinary least squares estimator, $b^{\bigstar} = b^{o}$. And the shrinkage estimator corresponding to $\delta_1 = \ldots = \delta_R = 0$ ( i.e. $\Delta = 0$ ) is $b^{\bigstar} = \vec{0}$. When there is no restriction on shrinkage factors other than $0 \leq \delta_j \leq 1$ for $1 \leq j \leq R$, the set of all possible generalized shrinkage estimators, $b^{\bigstar}$, constitutes the surface and interior of an R-dimensional hyper-rectangle with...

– each of its edges parallel to a corresponding principal axis of the given regressors,
– one of its $2^R$ vertices at the least squares estimate, $b^o$, and
– the "diagonally opposite" vertex at the shrinkage origin, which is usually $\vec{0}$ .

The above generalized shrinkage "geometry" is illustrated below, in Figure 3.1, for the P=R=2 dimensional case described in the simple N = 10 observation numerical example discussed in Sections §2.3, §2.6 and §2.10 of Chapter 2.

## Figure 3.1 Two-Dimensional Generalized Shrinkage Rectangle




Again, all points either on the boundary or in the interior of the rectangle of Figure 3.1 are generalized shrinkage estimators as defined by equation { 3.1 }.

## 3.1 Moments of Generalized Shrinkage Estimators

Generalized shrinkage estimators, { 3.1 }, can be linear estimators of $\beta$ for the fixed-effect model of equation { 2.1 } (or { 2.3 }) of Chapter 2. But they are linear only in cases where all

R of the generalized shrinkage factors, $\delta_1, \ldots, \delta_R$, are non-stochastic given X. In these special cases, the conditional expected value of $b^{\bigstar}$ will be

$$E(b^{\bigstar} \mid X) = G \Delta \gamma . \qquad \{ 3.3 \}$$

These shrinkage estimators, $b^{\bigstar}$, are generally "biased" estimators. After all, the shrinkage expectation vector is $G \Delta \gamma$, and this vector is generally $\neq \beta = G \gamma$ in cases where $\Delta \neq I$. The bias in $b^{\bigstar}$, namely $G (I - \Delta) \gamma$, is usually unknown in practical applications because, just like $\beta$ itself, the true $\gamma = G^T \beta$ components would also be unknown.

Similarly, the conditional variance matrix of $b^{\bigstar}$ for non-stochastic shrinkage in a fixed-effect model is

$$V(b^{\bigstar} \mid X) = \sigma^2 \cdot G \Delta^2 \Lambda^{-1} G^T . \qquad \{ 3.4 \}$$

Expression { 3.4 } is certainly not a universally valid variance formula; it simply does not apply when the shrinkage factor matrix, $\Delta$, is actually stochastic given X. And uncertainty about an appropriate form and extent for shrinkage occurs in most (if not all) practical applications! After all, a shrinkage practitioner generally chooses his/her desired shrinkage "pattern" only after examining several functions (statistics) that depend upon the observed response vector, y. For example, the practitioner may make his/her choice only after visually examining shrinkage TRACE displays ...looking for shrinkage that will "stabilize" coefficients and/or change the numerical signs of fitted coefficient estimates! Exact moment formulae for the resulting NONLINEAR shrinkage estimators frequently cannot be computed. In fact, it is usually difficult to conjecture whether equation {3.4} is even approximately "correct" in situations where one's choice of $\Delta$ shrinkage factors ends up being stochastic. However, nothing less than the above sorts of visualization strategies/tactics give sufficient insights to make "realistic" shrinkage decisions in almost all practical shrinkage-regression applications.

## 3.2 Shrinkage Inflation of the Residual Mean Square

The squared length of the residual vector corresponding to a shrinkage estimator $b^{\bigstar} = G \Delta \gamma$ is used to define the corresponding "inflated" residual mean square, $RMS^{\bigstar}$, as follows...

$$\begin{aligned} RMS^{\bigstar} &= [\,(y - X b^{\bigstar})^T (y - X b^{\bigstar})\,] / (N - R - 1), \qquad \{ 3.5 \} \\ &= [\; y^T (I - H \Delta H^T)^2 y \;] / (N - R - 1), \\ &= [\; y^T (I - H H^T) y + y^T H (I - \Delta)^2 H^T y \;] / (N - R - 1), \\ &= RMS^{o} + [\, y^T y / (N - R - 1) \,] \cdot r^T (I - \Delta)^2 r , \end{aligned}$$

where $RMS^{o} = s^2 = y^T (I - H H^T) y / (N - R - 1)$ is the least-squares residual mean square for error. Note that the $(N - R - 1)$ factor in the denominator of { 3.5 } makes $RMS^{o}$ an unbiased estimator of $\sigma^2$ in equations { 2.2 } and { 2.4 } under normal distribution theory, Johnson and Kotz(1970), equation (10), page 168. Furthermore, by the "Principle of Least Squares," $RMS^{o}$ is (essentially by its very definition) the minimum residual-mean-square. In

fact, the shrinkage residual-mean-square, RMS$^{\bigstar}$ of { 3.5 }, is usually a fairly "uninteresting" statistic that may grossly overestimate the true $\sigma^2$.

## 3.3 The Hoerl-Kennard "Ordinary" Ridge Family

The ridge estimators originally proposed by Hoerl and Kennard (1970a,b) are of the highly specialized form:

$$b^{\bigstar} \quad = \quad ( X^T X + k \cdot I )^{-1} X^T y . \qquad \{ 3.6 \}$$

However, simple matrix-algebraic manipulations, using the singular value decomposition of X in equation { 2.8 }, show that these estimators can be rewritten as...

$$\begin{aligned} b^{\bigstar} \quad &= \quad [ G ( \Lambda + k \cdot I ) G^T ]^{-1} G \Lambda^{1/2} H^T y , \\ &= \quad G ( \Lambda + k \cdot I )^{-1} \Lambda^{1/2} H^T y , \\ &= \quad G [ ( \Lambda + k \cdot I )^{-1} \Lambda ] \Lambda^{-1/2} H^T y , \\ &= \quad G \Delta c \end{aligned}$$

for $\Delta = ( \Lambda + k \cdot I )^{-1} \Lambda$ and $k \geq 0$. Equivalently, the corresponding generalized shrinkage $\delta$-factors are...

$$\delta_j \quad = \quad \lambda_j / ( \lambda_j + k ) \quad \text{for } 1 \leq j \leq R.$$

The shrinkage family of equation { 3.6 } is easily derived as a solution to either of two optimization problems:

(1) What is the locus of the most-likely $\beta$ estimate vectors for each given length?

and

(2) What is the locus of shortest $\beta$ estimate vectors for each given likelihood?

For example, the Lagrange equation with multiplier k for maximizing the likelihood that a vector of values b is $\beta$ (i.e. minimizing the corresponding residual sum-of-squares) subject to the restriction that the squared length of b is $b^T b = C^2$ can be written as

$$\Psi( b ) = ( y - Xb )^T ( y - Xb ) + k \cdot ( b^T b - C^2 ) ,$$

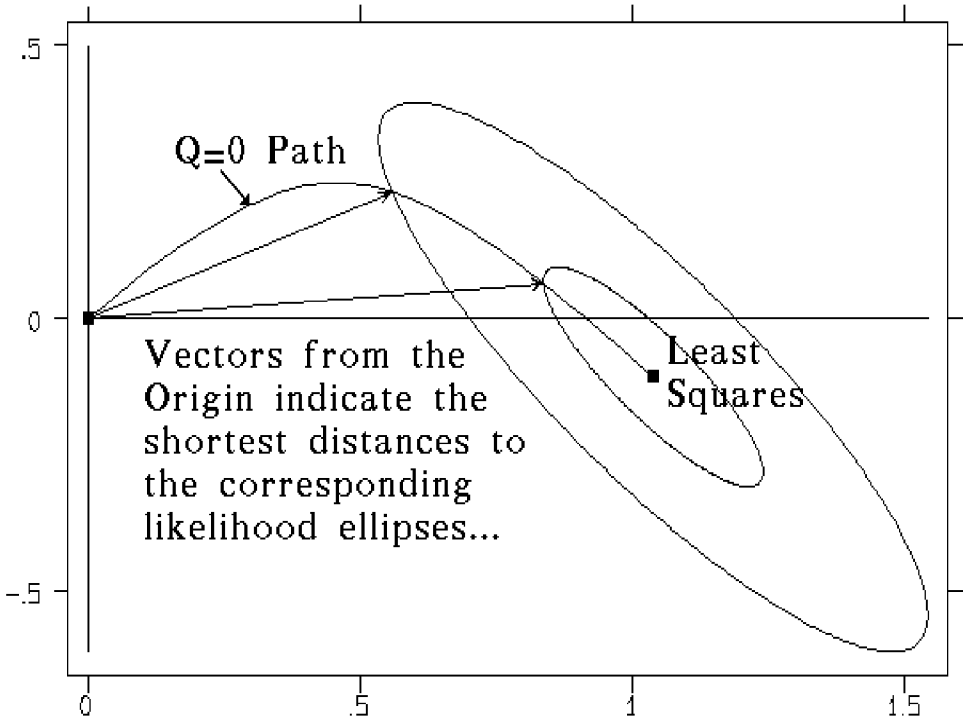
where the X matrix and the y vector have been "centered" as in equations { 2.3 } and { 2.4 }. Equating $\partial \Psi / \partial b = 2 \cdot ( X^T X b + k \cdot b - X^T y )$ to $\vec{0}$ then yields equation { 3.6 }. The "analytical geometry" of this situation is illustrated in Figure 3.2, below, for the simple $P = R = 2$ dimensional numerical example of Figure 3.1 and Chapter 2. The Hoerl-Kennard "ordinary" ridge shrinkage path is labeled "Q = 0" in Figure 3.2 for consistency with the notation to be introduced in Section §3.3, below.

Note also that equation { 3.6 } offers direct, intuitive appeal whenever a regression problem is ill-conditioned! Namely, it seems to be almost "obvious" that adding small, positive values (

$k \cdot I$ ) to the diagonal of $X^T X$ will make any nearly singular regressor inner-products matrix "easier" to invert; see Piegorsch and Casella(1989) for information on this and other early motivations for "ridge" regression terminology.

Equation { 3.6 } might seem to suggest that the ( $X^T X + k \cdot I$ ) matrix be re-inverted each time the numerical value of k is changed. In stark contrast, Equation { 3.1 }, which (as we have already seen) includes { 3.6 } as a special case, offers distinct computational advantages! Having once computed the least squares decomposition

**Figure 3.2 The Hoerl-Kennard Shrinkage Path**




$b^o = G c$ of { 2.14 }, equation { 3.1 } suggests that shrinkage estimates be calculated by simple matrix and vector multiplications, as $b^\star = G \Delta c$ . Meanwhile, the great "intuitive" advantage of equation { 3.1 } is that $b^\star = G \Delta c$ is immediately seen as resulting from different amounts of shrinkage along each principal axis of regressors. Finally, the "ordinary" ridge formula $\delta_j = \lambda_j / ( \lambda_j + k ) = 1 / ( 1 + k/\lambda_j )$ is seen to have an additional, intuitive appeal. Namely, for each fixed k value, greater shrinkage (a smaller $\delta_j$ factor) is applied to the least squares components corresponding to the smallest "spreads" (smallest $\lambda_j$ values) in the given regressor coordinates.

## 3.4 The Two-Parameter Generalized Ridge Family

Unlike the totally unrestricted approach of { 3.1 }, the shrinkage factors $(\delta_1, \ldots, \delta_R)$ for most "families" of interest are functions of at most two parameters. The pair of shrinkage "hyper-parameters" we discuss below (MCAL and QPAR) are...

(i) not only adequate to control both the form (or general shape) and the extent of shrinkage,

(ii) but also general enough to include, as special cases, the shrinkage families considered by the vast majority of authors who have published descriptions of generalized shrinkage strategy/tactics.

First of all, as shrinkage occurs, $b^{\bigstar}$ generally moves away from $b^{o}$ and toward $\vec{0}$. The primary shrinkage parameter can thus be taken to be:

$$\begin{aligned} \text{MCAL} &= \text{M}, \quad \text{the "multicollinearity allowance" parameter that controls the extent of shrinkage,} \\ &= R - \delta_1 - \delta_2 - \ldots - \delta_R. \end{aligned} \qquad \{ 3.7 \}$$

Shrinkage of $b^{\bigstar}$ from $b^{o}$ towards $\vec{0}$ follows a "path" whose general shape can also be controlled by the regression practitioner. Our secondary shrinkage parameter will be denoted by:

$$\text{QPAR} = \text{Q}, \quad \text{the "shape" parameter that controls the curvature of the shrinkage path through regression coefficient likelihood space.}$$

Specifically, the primary 2-parameter functional form for shrinkage factors considered here will be:

$$\begin{aligned} \delta_j &= \lambda_j \; / \; [\; \lambda_j + \text{Konst.} \; \lambda_j^{Q} \;] \\ &= 1 \; / \; [\; 1 + \text{Konst.} \; \lambda_j^{Q-1} \;], \end{aligned} \qquad \{ 3.8 \}$$

where the constant, Konst, in { 3.8 } is chosen to provide any specified extent of shrinkage, as quantified by $M = R - \delta_1 - \delta_2 - \ldots - \delta_R$ of { 3.7 }. One interesting implication of this 2-parameter family is this: Unless $Q = 1$, Konst. $= 0$, or Konst. $= +\infty$, a pair of delta shrinkage factors can be equal, $\delta_i = \delta_j$, if and only if their corresponding regressor eigenvalue (sum-of-squares) are also equal, $\lambda_i = \lambda_j$. Because regressor eigenvalues are "usually" distinct (except, say, for designed experiments), the shrinkage estimators in the 2-parameter family of { 3.8 } "usually" do a different amount of shrinkage along each principal axis of regressors.

The two-parameter family of { 3.8 } was apparently first published in Goldstein and Smith (1974), equation (13), where our Q parameter is $1 - m$ in their notation (and m was explicitly

restricted to integer values.) Equation (14) of Goldstein and Smith(1974) and our { 3.8 } can both be rewritten as:

$$b^{\bigstar} = [\, X^T X + k \cdot (\, X^T X \,)^Q \,]^{-1} X^T y \,. \qquad \{ 3.9 \}$$

Restriction of Q to integer values is not necessary, of course. I, personally, have found a somewhat finer mesh of Q-shapes (including at least half-integer values) useful in practical applications. On the other hand, it would seem rather silly, at least to me, to search for a "best" Q-shape to within, say, three or even more decimal places!

Considerable confusion about the 2-parameter family of { 3.8 } has been voiced in statistical literature. For example, Hoerl and Kennard(1975) point out that this family (where our Q is $-q$ in their notation) was proposed by R. W. Somers in a 1964 presentation at an A.I.Ch.E. Symposium in Memphis, Tennessee. Hoerl and Kennard(1975) suggest the restriction that Q be $\leq 0$ and integral, and they also express doubts, without really saying why, that this generalization of their original $Q = 0$ proposal will be of any value in practical applications. Draper and Van Nostrand(1979) only heighten this confusion by failing to recognize the relationship between our equations { 3.8 } and { 3.9 } and their equations (3.10) and (3.29). And they repeat the $Q \leq 0$ restriction. We will demonstrate below [and in our shrinkage regression case studies] that the full 2-parameter family of { 3.8 } is not only extremely versatile but also quite useful in actual practice.

To illustrate generalized shrinkage paths of different Q – shapes in Figure 3.3 below, we again return to the P=R=2 dimensional, $N = 10$ observation numerical example discussed in Chapter 2 and used in Figures 3.1 and 3.2. Besides the Hoerl-Kennard ($Q = 0$) path displayed in Figure 3.2, Figure 3.3 also explicitly shows the paths for the $Q = +2$, $Q = +1$, and $Q = -1$ shapes. Furthermore, the two upper edges of the generalized shrinkage rectangle represent the limiting Q-shape path as Q approaches $-\infty$, while the two lower edges represent the limiting Q-shape as Q approaches $+\infty$.

The following classification of numerical values for the QPAR = Q path shape parameter of { 3.8 } and { 3.9 } is of at least historical interest...

(a) QPAR > 1.0 yields what are commonly called "increasing" $\delta$ factors.

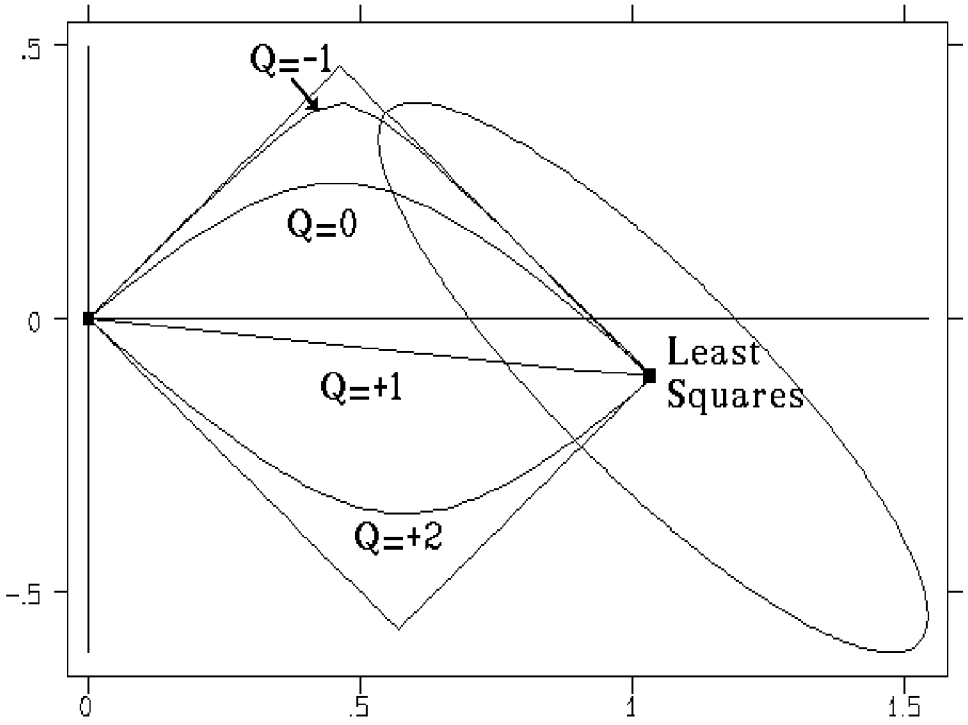
Technically, QPAR > 1.0 implies only that $\delta_1 \leq \ldots \leq \delta_R$ . But, again, two $\delta$ factors can be equal only when their corresponding eigenvalues are equal. Thus QPAR > 1.0 frequently implies $\delta_1 < \ldots < \delta_R$ .

These shrinkage patterns strike me as being somewhat counterintuitive; they actually shrink the relatively precise components of $b^o$ more than its relatively imprecise components. Authors like Thisted(1976), Strawderman(1978) and Casella(1980, 1985) underscore the following paradox: minimax ridge estimators tend to concentrate shrinkage along whichever axes the practitioner regards as LEAST important for improvements in mean-squared-error. Minimax shrinkage patterns with QPAR > 1.0

(shrinkage primarily along major axes) can thus result when the user thinks that he/she has emphasized risk minimization along minor axes!

Strawderman(1978) points out that the two loss functions most commonly used in statistical decision theory correspond to QPAR = 2 and QPAR = 1; for more details, see Section §9.7.

**Figure 3.3 Several Q-Shape Shrinkage Patterns**




(b) QPAR = 1.0 yields $\delta$ factors that are all equal, $\delta_1 = \ldots = \delta_R$ .

BLUP estimates for "balanced" designs as well as Stein-type estimates are usually of this special form. Mayer and Willke(1972) called them "shrunken" estimators, but the terminology "uniformly shrunken" estimators would, of course, be more descriptive.

The coefficient trace display is visually uninteresting when the Q – shape is + 1. This trace then consists of P straight lines, each running from a least squares estimate at M = 0 directly to ZERO at M = P. In other words, the relative magnitudes of the elements of $b^{\bigstar}$ are always exactly identical to those of $b^{o}$ in the QPAR = + 1 special case.

Similarly, the trace of shrinkage factors is visually uninteresting when QPAR=1; all P shrinkage factors plot right on top of each other! All shrinkage factors fall on the straight line from $\delta_j = 1$ at M = 0 directly to $\delta_j = 0$ at M = R.

On the other hand, trace plots of estimated mean-squared-error, excess eigenvalues, and the inferior direction can still lead to new insights even when the QPAR shape is $+1$.

(c) QPAR < 1.0 yields what are commonly called "declining $\delta$ factors."

Technically, QPAR < 1.0 implies only that $\delta_1 \geq \delta_2 \geq \ldots \geq \delta_R$ , but (again) two $\delta$ factors can be equal only when the corresponding eigenvalues are equal. Thus QPAR < 1.0 frequently implies $\delta_1 > \ldots > \delta_R$ .

QPAR = 0.0 is, of course, the original, "ordinary" form of ridge regression suggested by Hoerl and Kennard(1970a,b); see equation { 3.6 }.

QPAR = $-1.0$ is an option that yield's delta's which decline more markedly than in the original Hoerl-Kennard formulation; see also Crone(1972) and remarks by Goldstein and Smith (1974) on increasing sensitivity to the "eigenvalue spectrum." Q = $-1$ yield $\delta$ factors of the form...

$$\delta_j = 1 \; / \; [\; 1 + \text{Konst.} \; \lambda_j^{-2} \;] \qquad \{ 3.10 \}$$

(d) The limit as QPAR approaches $-\infty$ (minus infinity) yields "principal-components regression" estimates.

When the ordered eigenvalues of regressor spread are strictly decreasing ( $\lambda_1 > \lambda_2 > \ldots > \lambda_R$ without ties), values of QPAR that are negative and large yield shrinkage patterns that are very close, numerically, to what is commonly called Principal-Components Regression; see Kendall(1957) and Massy(1965), method "a", page 241. This "extreme" shrinkage pattern corresponds to moving along a certain chain of edges of the shrinkage-factor hypercube leading from the $\Delta = I$ vertex to the diagonally opposite $\Delta = 0$ vertex. Equivalently, this is the special case where shrinkage factors are non-increasing ( $\delta_1 \geq \delta_2 \geq \ldots \geq \delta_R$ ), and all shrinkage factors, except at most one, are always either 1 or 0 at each point along the path. Marquardt(1970) terms these same estimates "fractional-rank" or "generalized inverse" estimates.

The coefficient TRACE for this shrinkage path consists of broken but connected line segments, with break-points at every integer value of MCAL = M.

For completeness, we note that a 2-parameter family of shrinkage factors that can be quite different from those of { 3.8 } is given by:

$$\delta_j = \min( 1, \text{Konst} \cdot \lambda_j^{Q} ) \qquad \{ 3.11 \}$$

where the constant, Konst, in { 3.11 } is chosen to provide any specified extent of shrinkage, again quantified by $M = R - \delta_1 - \delta_2 - \ldots - \delta_R$ of { 3.7 }. For a fixed power Q, shrinkage starts in { 3.11 } with any sufficiently large Konst value so that all $\delta$ factors will equal 1. This initial Konst value can be taken to be max( $\lambda_1^{-Q}$, $\lambda_R^{-Q}$ ). One or more $\delta$ factor then starts decreasing as Konst decreases, but at least one $\delta$ factor remains fixed at 1 until Konst drops below min( $\lambda_1^{-Q}$, $\lambda_R^{-Q}$ ). From this point onward, all $\delta$-factors decrease at the same rate, and all shrunken coefficients have fixed relative magnitudes, converging to zero at Konst = 0. Note the following special path-shapes in this secondary 2-parameter family...

Q = 0.0 for uniform (Stein-like) shrinkage,

Q = 0.5 leads to shrunken coefficients that become both uncorrelated and homoscedastic, as in the equity estimator of Krishnamurthi and Rangaswamy(1987,1989),

and

Q = 1.0 leads to shrunken coefficients that approach the exact same relative magnitudes as the marginal inner-products vector (and are thus guaranteed to have no coefficients with "wrong" signs), as in Obenchain(1978) and equation { 4.17 }.

The secondary 2-parameter family of { 3.11 } strikes me as being somewhat less versatile and somewhat more cumbersome to apply than the primary family of { 3.8 }. For example, closed-form expressions for maximum likelihood estimates within the primary family will be introduced in Chapter 5, and these statistics greatly facilitate choice of Q-shape (as well as shrinkage extent) within the primary family. By way of contrast, no such closed-form expressions are available for the secondary family; choice of Q-shape is thus left either to personal preference or to relatively tedious numerical searches along a variety of different path shapes.

I certainly hope that readers will not be too confused by use of the SAME symbols (Q and Konst) and terminology for BOTH of the 2-parameter families considered here in Section §3.4. These two families are really quite different! For example, uniform shrinkage is Q=1 in the primary family but Q=0 in the secondary family. Similarly, ridge shrinkage requires the Konst to increase in the primary family but to decrease in the secondary family! Enough said?

## 3.5 The Implicit Intercept for Shrinkage

Most of our attention, so far, has been focused upon the generalized shrinkage estimates, $b^{\bigstar}$, for the P elements of $\beta$ corresponding to non-constant regressor variables. The resulting "implicit" estimate for the intercept term, $\mu$, is $\bar{y} - \bar{x}^T b^{\bigstar}$ for each $b^{\bigstar}$. (We will see in Chapter 5,

equation { 5.3 }, that this is the Normal-distribution-theory maximum likelihood estimate of $\mu$ corresponding to any $b^{\bigstar}$ estimate of $\beta$.) Note that this intercept estimate will usually approach $\bar{y}$, NOT zero, as the shrinkage $\delta$-factors approach zero. In other words, shrinkage of regression coefficients to zero can also be visualized as simply a rotation of the fitted regression hyperplane about the $( \bar{x}^T, \bar{y} )$ point so that it becomes more horizontal (or "flat") along all P regressor coordinate axes. Generally speaking, the point-of-view adopted here is that the intercept estimate changes in shrinkage regression only because the estimates of the $\beta$ coefficients are changing. This perspective is illustrated in Figure 3.4 below.

If a formula like { 3.6 } were used to estimate $( \mu, \beta^T )$ without first "centering" either the response y vector or the augmented regressor ( 1, X ) matrix, namely

$$\begin{pmatrix} \mu^{\bigstar} \\ b^{\bigstar} \end{pmatrix} = \left\{ \begin{bmatrix} N & 1^T X \\ X^T 1 & X^T X \end{bmatrix} + k \cdot I \right\}^{-1} \begin{pmatrix} 1^T y \\ X^T y \end{pmatrix},$$

then the $\mu^{\bigstar}$ intercept term would be shrunk to zero just like the $b^{\bigstar}$ coefficient estimates. This sort of situation is illustrated in Figure 3.5, above. Note that shrinkage of this "nonstandard" form can quickly become drastic in the sense that the fit can "miss" all of the data! In other words, all observed data points ultimately end up on the same side of the fitted, shrinkage hyperplane (i.e. on the same side of the fitted line in the $P = R = 1$ case shown in Figure 3.5).

**Figure 3.4 The Implicit Shrinkage Intercept**

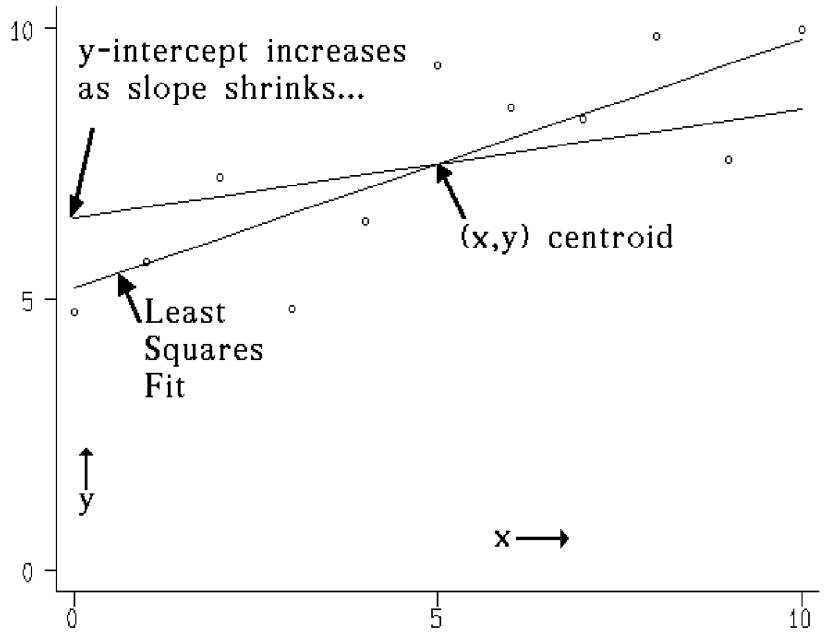
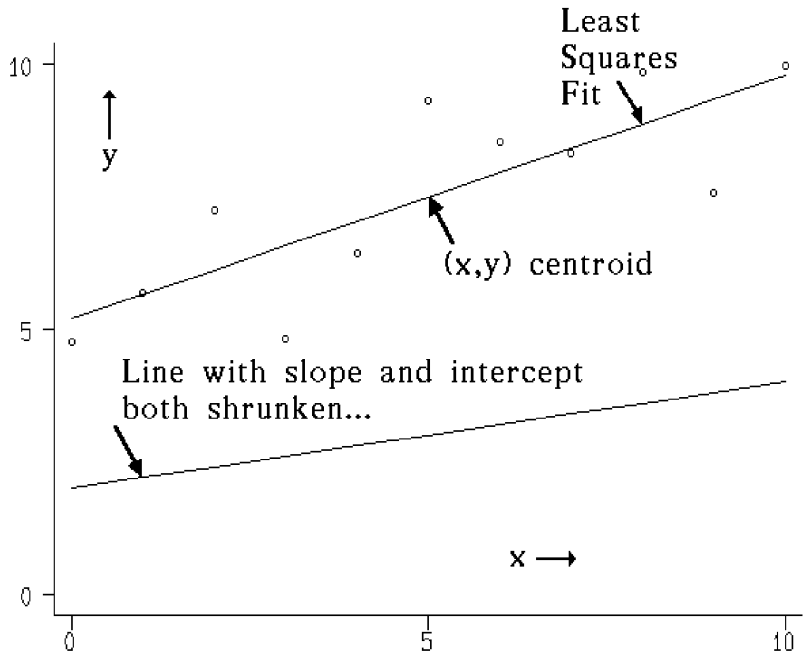



**Figure 3.5 Shrinking Both Intercept and Slope**



## 3.6 Models Without an Intercept Term

Models without an intercept, the $1 \cdot \mu$ term of formula { 2.1 }, cannot be analyzed "exactly" using "centered" variables, as in formulas { 2.3 } and { 2.4 }. (Technically, even when the explicit $1 \cdot \mu$ term is absent, the model still actually includes an intercept whenever 1 lies within the column space of the non-constant regressors X matrix BEFORE it has been "centered.") Models without an intercept restrict the regression fit to pass through $y = 0$ at $x = \vec{0}$ instead of through $y = \overline{y}$ at $x = \overline{x}$. In fact, centering can be visualized as a convenient mechanism for assuring that the ( $\overline{x}^T$, $\overline{y}$ ) pivot point is translated so as to coincide with ( $\vec{0}^T$, 0 ). One looses only a single degree-of-freedom for error in estimating $\beta$ by pre-multiplying both the response y vector and the non-constant regressor X matrix of a model with no intercept on the left by the "centering" projection matrix ( $I - 1\,1^T / N$ ). And we argue below that, in essence, using this wrong (centered) model can actually make sense in certain practical applications.

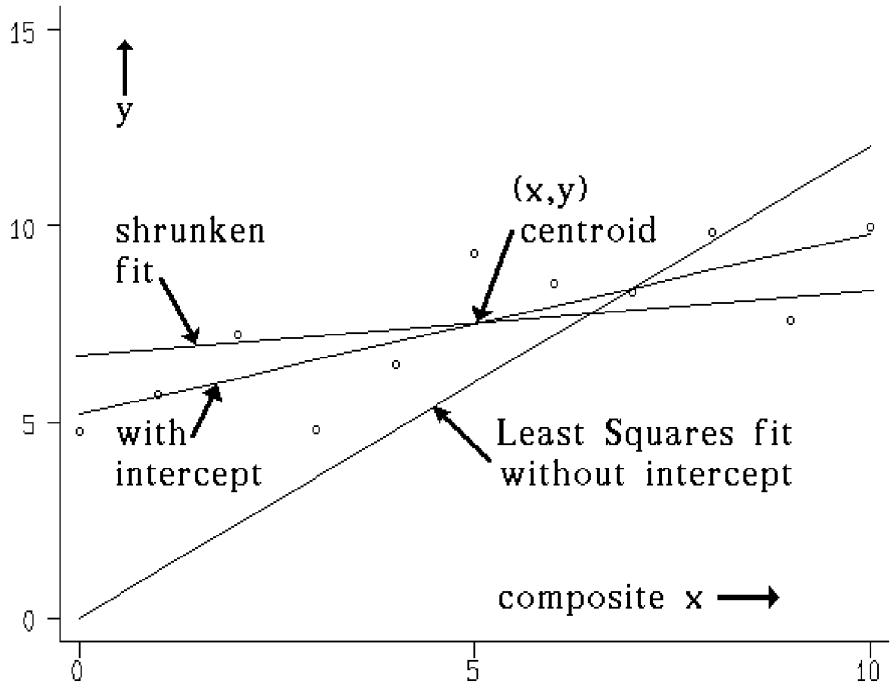
The only information not being used in the centered version of a model with no intercept is that expressed by the restriction $\overline{y} = \overline{x}^T\beta$; in models with an intercept, the difference $\overline{y} - \overline{x}^T\beta$ is simply the "implicit" intercept at $x = \vec{0}$, as described in Section §3.5 above. In fact, it is tempting to proceed as if $\overline{y} = \overline{x}^T\beta$ is merely a restriction on the length of the regression coefficient vector, $\beta$, that doesn't need to be addressed until we have reached the final, visual re-regression (VRR) phase of our analyses.

On the other hand, Figure 3.6 below illustrates a case where the information we ignored from the single degree-of-freedom for "no intercept" turns out to be rather traumatic once we have reached the final, VRR phase of our analysis. Here, both the shrinkage regression fit and the least-squares fit for a model with an intercept term are represented by a pair of lines which pass through $\overline{y}$ at $\overline{x}^T\beta^\bigstar$, where $\beta^\bigstar$ is the unit vector parallel to our favorite shrinkage regression estimator of $\beta$ coefficients for the non-constant regressors. Unfortunately, neither of these two lines comes anywhere near to $y = 0$ at $x^T\beta^\bigstar = 0$. And yet only lines that actually do pass through $y = 0$ at $x^T\beta^\bigstar = 0$ are even candidates now that we have reached the final VRR phase of our analysis for a no-intercept model. Note that it would be truly unreasonable to simply translate (by moving parallel to itself) either one of these two fitted lines (down) so that it would pass through the ( 0, 0 ) origin; the fitted line would then again totally "miss" the data! My personal reaction in any sort of situation like that of Figure 3.6 would be that the available data cast very serious doubt on the cogency of a no-intercept model. After all, the least-squares line through ( 0, 0 ) would then yield mostly positive residuals to the left of the composite regressor mean value, $\overline{x}^T\beta^\bigstar$, and mostly negative residuals to the right of this same point.

On the other hand, the situation depicted in Figure 3.6 is somewhat pathological and, in many cases, the no-intercept least-squares fit will be a viable reference line for consideration during VRR. The equation for this least-squares fit is derived as follows: Let the $N \times 1$ vector of un-centered response values be $y^* = y + \overline{y}^* 1$, where y is centered; let the $N \times P$ matrix of un-centered, non-constant regressor coordinates be $X^* = X + 1 \cdot \overline{x}^{*T}$, where X is centered; and consider only rescaled estimates of the original shrinkage $b^\bigstar$ of the form $\widehat{\beta} = f \cdot \beta^\bigstar$. The sum-of-squares to be minimized is then $(y^* - X^*\widehat{\beta})^T(y^* - X^*\widehat{\beta})$, and the best rescaling is

$$f \;=\; \frac{y^{*T}X^{*}\beta^{\bigstar}}{\beta^{\bigstar T}X^{*T}X^{*}\beta^{\bigstar}} \;=\; \frac{(y^{T}X+N\cdot\bar{y}^{*}\cdot\bar{x}^{*T})\beta^{\bigstar}}{\beta^{\bigstar T}(X^{T}X+N\cdot\bar{x}^{*}\bar{x}^{*T})\beta^{\bigstar}}\;. \qquad \{\,3.12\,\}$$

**Figure 3.6 Least Squares and Shrunken Fits For a Model with No Intercept**




## 3.7 Shrinkage Residual Analyses

The residual vector, $r^{\bigstar}$ , corresponding to the generalized shrinkage estimator, $b^{\bigstar}$ , of equation { 3.1 } is

$$\begin{aligned} r^{\bigstar} \;&=\; y \,-\, X\,b^{\bigstar}, \\ &=\; (\,I \,-\, H\,\Delta\,H^{T}\,)\,y\,, \\ &=\; (\,I \,-\, 1\,1^{T}/N \,-\, H\,\Delta\,H^{T}\,)\,y\,, \end{aligned} \qquad \{\,3.13\,\}$$

where that last expression applies even when the response vector, y , hasn't been centered (so that $1^{T}\,y = 0$.)

Warning about possible confusions in notation: The $R \times 1$ vector of principal correlations of equation { 2.16 } is denoted by the symbol r , i.e. with no superscript. And, when individual elements of a correlation vector are referenced, they will have two subscripts; either $r_{yj}$ for principal correlations or $r_{yx_j}$ for marginal correlations, respectively. In contrast, the symbol $r^{\bigstar}$ (with a star superscript) represents an $N \times 1$ vector of fitted shrinkage-regression residuals. Note also that elements of the $r^{\bigstar}$ residual vector would be written with only one subscript ($r_i^{\bigstar}$ for the i-th observation.)

The shrinkage-regression residual vector can have a non-zero expected value even when $E(y \mid X) = X\beta$ of equation { 2.3 } is a correct expectation model because shrinkage estimators are usually biased estimators:

$$E(r^{\bigstar} \mid X) = (I - H \Delta H) X \beta = H (I - \Delta) \Lambda^{1/2} \gamma . \qquad \{ 3.14 \}$$

One sufficient condition for $E(r^{\bigstar} \mid X) = 0$ when the model is correct is that $\Delta = I$ ; after all, absolutely no shrinkage results in this extreme case (where $b^{\bigstar} = b^{o}$ and $r^{\bigstar} = r^{o}$ of ordinary least squares.)

The conditional variance-covariance matrix of the shrinkage-regression residual vector, given the observed regressor coordinates and under the assumption that $V(y \mid X) = \sigma^2 \cdot (I - 1\,1^T / N)$ of equation { 2.4 } is a correct dispersion model, is of the general form:

$$\begin{aligned} V(r^{\bigstar} \mid X) &= \sigma^2 \cdot (I - 1\,1^T / N - H \Delta H^T)^2 , \\ &= \sigma^2 \cdot (I - 1\,1^T / N - H H^T + H (I - \Delta)^2 H^T) . \\ &= V(r^{o} \mid X) + \sigma^2 \cdot H (I - \Delta)^2 H^T . \end{aligned} \qquad \{ 3.15 \}$$

### 3.7.1 Leverage Modifications Resulting from Shrinkage.

With $x_i^T$ denoting the i-th row of the given regressor-coordinate X matrix, the shrinkage-regression prediction of the expected response at $x_i^T$ would be

$$\text{estimated } E(y_i \mid x_i) = x_i^T b^{\bigstar} = h_i^T \Delta r \cdot \sqrt{y^T y} , \qquad \{ 3.16 \}$$

where $h_i^T$ is the i-th row of the standard principal coordinates matrix, H of { 2.8 }, and r is the vector of principal correlations of equation { 2.16 }. The estimated variance of this prediction is

$$\text{estimated } V(y_i \mid x_i) = x_i^T V(b^{\bigstar} \mid X) x_i = \sigma^2 \cdot h_i^T \Delta^2 h_i , \qquad \{ 3.17 \}$$

where $\sigma^2$ is estimated by the least-squares residual-mean-square, $s^2 = RMS^{o}$.

Again, we define the leverage of the i-th observation on the regression, as in equation { 2.50 }, to be

$$\Lambda_i = \frac{\text{Predictive Variance}}{\text{Residual Variance}} = \frac{h_i^T \Delta^2 h_i}{[\,(N-1)/N - h_i^T h_i + h_i^T (I-\Delta)^2 h_i\,]} . \qquad \{ 3.18 \}$$

It is clear from equation { 3.18 } that shrinkage can only reduce the leverage of every regressor combination! After all, the numerator predictive variance is maximized and the denominator residual variance is minimized at the ordinary least squares solution; $\max \Lambda_i = h_i^T h_i / [\,(N-1)/N - h_i^T h_i\,]$ is achieved at $\Delta = I$.

When $h_i^T h_i$ is relatively large for the i-th observation, this means that $x_i^T$ (the i-th row of X ) is rather large as measured in the metric of $(X^T X)^{-1}$ and, thus, that the i-th regressor combination is relatively remote from the centroid of (possibly highly ill-conditioned) regressor coordinates. This remoteness gets translated by the least-squares fitting algorithm into a corresponding small residual variance, implying that the least-squares fit will be pulled relatively close to the corresponding observed response value, $y_i$.

When the shrinkage pattern is not uniform, some elements of $\Delta$ will be decreasing much more rapidly than others. For example, if $h_i^T = (\, h_{i1}, h_{i2}, \ldots h_{iR}\,)$ has some relatively large trailing coordinates [i.e. $|\, h_{iR}\,|, |\, h_{i(R-1)}\,|, \ldots$ are large relative to $|\, h_{i1}\,|, |\, h_{i2}\,|, \ldots$] and the shrinkage pattern utilizes declining deltas [ $\delta_1 > \delta_2 > \ldots > \delta_R$ ], then the leverage of that regressor combination will decrease very quickly with shrinkage. After all, such an $h_i^T = (\, h_{i1}, h_{i2}, \ldots h_{iR}\,)$ gets most of its leverage from the minor-principal-axis dimensions that shrinkage is systematically de-emphasizing.

### 3.7.2 Standardized/Studentized Shrinkage Residuals.

The i-th residual is standardized by dividing it by the "usual" estimate of its standard deviation, the square root of the i-th diagonal element of { 3.15 } with $\sigma^2$ estimated by $s^2$ of { 2.35 } :

$$r_i^{\bigstar s} = \frac{r_i^{\bigstar}}{s \cdot \sqrt{(N-1)/N - h_i^T (\,2\Delta - \Delta^2\,) h_i}} . \qquad \{ 3.19 \}$$

These standardized residuals do not follow Student's-t distribution under normal theory because $r_i^{\bigstar}$ and s are not statistically independent. As in equation { 2.46 } , the estimator of $\sigma^2$ that is independent of $r_i^{\bigstar}$ is $s^2_{(-i)}$ of

$$(\, N - P - 2\,) \cdot s^2_{(-i)} = (\, N - P - 1\,)\, s^2 - (\, r_i^o\,)^2 / [(N-1)/N - h_i^T h_i\,] . \qquad \{ 3.20 \}$$

The i-th shrinkage residual is thus studentized as follows:

$$t_i^{\bigstar} = \frac{r_i^{\bigstar}}{s_{(-i)} \cdot \sqrt{(N-1)/N - h_i^T (2\Delta - \Delta^2) h_i}}, \qquad \{ 3.21 \}$$

$$= r_i^{\bigstar s} \cdot \sqrt{\frac{N-P-2}{[N-P-1-(r_i^{\bigstar s})^2]}}$$

The results derived here in Section §3.5 on analysis of shrinkage residuals can be summarized as follows:

> Shrinkage reduces the overall leverage of every regressor combination. But, when shrinkage is not uniform, the leverage of some observations may be reduced much more quickly than others. While shrinkage increases the average size of fitted residuals, some residuals may actually become smaller. In other words, shrinkage can change not only the relative magnitudes of fitted residuals but also of their standardized and studentized versions.

# References for Chapter Three


Casella, G. (1980). "Minimax ridge regression estimation." **Annals of Statistics** 8, 1036-1056.

Casella, G. (1985). "Condition numbers and minimax ridge-regression estimators." **Journal American Statistical Association** 80, 753-758.

Crone, L. (1972). "The singular value decomposition of matrices and cheap numerical filtering of systems of equations." **Journal Franklin Institute** 294, 133-136.

Draper, N. R. and Smith, H. (1981). **Applied Regression Analysis**, Second Edition. New York: John Wiley.

Draper, N. R. and Van Nostrand, R. C. (1979). "Ridge regression and James-Stein estimation: review and comments." **Technometrics**, 21, 451-466.

Hoerl, A. E. (1962). "Application of ridge analysis to regression problems." **Chemical Engineering Progress** 58, 54-59.

Hoerl, A. E. and Kennard, R. W. (1970a). "Ridge regression: biased estimation for nonorthogonal problems." **Technometrics** 12, 55-67.

Hoerl, A. E. and Kennard, R. W. (1970b). "Ridge regression: applications to nonorthogonal problems." **Technometrics** 12, 69-82.

Hoerl, A. E. and Kennard, R. W. (1975). "A note on a power generalization of ridge regression." **Technometrics**, 17, 269.

Goldstein, M. and Smith, A. F. M. (1974). "Ridge-type estimators for regression analysis." **Journal Royal Statistical Society**, B, 36, 284-291.

Johnson, N. L. and Kotz, S. (1970). **Distributions in Statistics: Continuous Univariate Distributions-1**. (Chapter 17, Gamma Distribution, including "Chi Square.") New York, John Wiley.

Krishnamurthi, L. and Rangaswamy, A. (1987). "The equity estimator for marketing research." **Marketing Science** 6, 336-357.

Krishnamurthi, L. and Rangaswamy, A. (1989). "Response function estimation using the equity estimator." Warton Working Paper 89-030R, University of Pennsylvania.

Lowerre, J. M. (1974). "On the mean square error of parameter estimates for some biased estimators." **Technometrics**, 16, 461-464.

Marquardt, D. W. (1970). "Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation." **Technometrics** 12, 591-612.

Massy, W. F. (1965). "Principal components regression in exploratory statistical research." **Journal American Statistical Association** 60, 234-256.

Mayer, L. S. and Wilkie, T. A. (1973). "On biased estimation in linear models." **Technometrics**, 15, 497-508.

Obenchain, R. L. (1975b). "Ridge analysis following a preliminary test of the shrunken hypothesis." **Technometrics**, 17, 431-441. (Discussion: McDonald, G. C., 443-445.)

Obenchain, R. L. (1978). "Good and optimal ridge estimators." **Annals of Statistics**, 6, 1111-1121.

Piegorsch, W. W. and Casella, G. (1989). "The early use of matrix diagonal increments in statistical problems." **Siam Review** 31, 428-434.

Rao, C. R. (1973). **Linear Statistical Inference and its Applications,** 2nd edition. New York: John Wiley & Sons.

Strawderman, W. E. (1978). "Minimax adaptive generalized ridge regression estimators." **Journal American Statistical Association** 73, 623-627.

Theil, H. (1963). "On the use of incomplete prior information in regression analysis." **Journal American Statistical Association** 58, 401-414.

Thisted, R. (1976). "Ridge regression, minimax estimation, and empirical bayes methods." **Technical Report No. 28, Division of Biostatistics**, Stanford University.